Available online at www.sciencedirect.com
SCIENCE DIRECT®
Chemometrics and Intelligent Laboratory Systems 70 (2004) 75–82
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Standard error of prediction in parallel factor analysis of three-way data ☆

Alejandro C. Olivieri[a,*], Nicolaas (Klaas) M. Faber[b,1]

[a] Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Suipacha 531, Rosario S2002LRK, Argentina
[b] Department of Production and Control Systems, ATO, P.O. Box 17, 6700-AA Wageningen, The Netherlands

Received 27 December 2002; received in revised form 11 July 2003; accepted 3 October 2003

Abstract

A simple approach is described to calculate sample-specific standard errors for the concentrations predicted by a three-way parallel factor (PARAFAC) analysis model. It involves a first-order error propagation equation in which the correct sensitivity and leverage values are introduced. A comparison is made with a related unidimensional partial least-squares (PLS) model, specifically as regards the required leverage values. Monte Carlo simulation results obtained by adding random noise to both concentrations and instrumental signals for theoretical binary mixtures are in good agreement with the proposed approach. An experimental multicomponent example was studied by a similar Monte Carlo approach, and the obtained standard errors are also in agreement with the calculated values. Implications concerning the limit of detection are discussed.

Keywords: Parallel factor (PARAFAC) analysis; Uncertainty propagation; Monte Carlo simulation; Standard error of prediction

1. Introduction

High-order arrays of data are particularly useful for the quantitative analysis of complex multicomponent samples. Specifically, three-way data following the trilinear model are gaining widespread analytical acceptance [1]. Interestingly, the decomposition of a three-dimensional array of data is often unique, allowing relative concentrations and spectral profiles of individual sample components to be extracted directly. Although there are many trilinear models, only the parallel factor (PARAFAC) analysis model ensures uniqueness in a straightforward way [2].

It is customary to report analytical results by including figures of merit such as concentration uncertainties and limits of application [3]. In the multivariate calibration context, average concentration errors may be estimated by studying a number of test samples with known analyte concentrations [4,5]. Sample-specific errors instead of average values can be assessed by replicate measurements, although Monte Carlo generated replications [6,7], simulations [8] and resampling methods [9] may also be useful. The most convenient approach seems to be the estimation of standard errors in predicted values with an uncertainty propagation equation. Usually, uncertainties are considered to be present in the calibration concentrations or reference values and also in the measured analytical signals (in the latter case for both calibration and unknown samples) [10].

Previous works have shown that uncertainty propagation works reasonably well in first-order multivariate calibration using partial least-squares (PLS) regression, principal component regression (PCR) and classical least-squares (CLS) analyses [9,10]. Recently, pre-processed multivariate calibration methods such as orthogonal signal correction (OSC) and net analyte pre-processing (NAP) [11] were also shown to lead to prediction errors, which are accounted for by simple error propagation theory. Likewise, reasonable uncertainty results were obtained for non-linear methods such as artificial neural networks (ANN) and non-linear PLS regression [12].

Several approaches have also been advanced in the context of three-way multivariate calibration using second-order data [13–17]. Sample-specific variance expressions

☆ Supplementary data associated with this article can be found in the online version, at doi:10.1016/j.chemolab.2003.10.005.

* Corresponding author. Tel.: +54-341-4372-704; fax: +54-341-4372704.
E-mail address: aolivier@fbioyf.unr.edu.ar (A.C. Olivieri).
[1] Current address: Chemometry Consultancy, Rubensstraat 7, 6717 VD Ede, The Netherlands.

0169-7439/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2003.10.005

Table 1
Stepwise description of calibration with three-way data using PARAFAC

Step	Operation[a]
1	Build an $I \times J \times K$ array with the I matrices for the I training samples
2	Decompose the array and obtain **A**, **B** and **C**
3	Identify the nth analyte of interest from **B** and **C** profiles
4	Calculate the relative concentrations in the unknown, $\mathbf{a}_{unk} = \mathbf{Z}^{+}\text{vec}(\mathbf{X}_{unk})$ and select the one of interest: $a_{n,unk}$
5	Regress the I elements of the column $\mathbf{a}_n$ against known standard concentrations $\mathbf{y}_{n,cal}$ for analyte n: $\mathbf{a}_n = k \times \mathbf{y}_{n,cal}$
6	Convert relative to absolute concentration of n in the unknown: $y_{n,unk} = a_{n,unk}/k$

[a] In step 4, vec($\mathbf{X}_{unk}$) represents the vector obtained by stringing out (unfolding) the unknown sample data matrix $\mathbf{X}_{unk}$ column-wise into a column vector. The matrix **Z** is given by $\mathbf{Z} = \mathbf{C}|\otimes|\mathbf{B}$, where the symbol $|\otimes|$ stands for the Khatri-Rao product [28], defined as follows: $\mathbf{C}|\otimes|\mathbf{B} = [\mathbf{c}_1 \otimes \mathbf{b}_1, \mathbf{c}_2 \otimes \mathbf{b}_2 \ldots \mathbf{c}_N \otimes \mathbf{b}_N] = [\text{vec}(\mathbf{b}_1\mathbf{c}_1^{T})\ \text{vec}(\mathbf{b}_2\mathbf{c}_2^{T}) \ldots \text{vec}(\mathbf{b}_N\mathbf{c}_N^{T})]$. The matrices **C** and **B** must have the same number of columns.

have been employed for the generalized rank annihilation method (GRAM) [15] and also for three-way partial least-squares (nPLS) [16]. As regards the highly useful PARAFAC model, work is progressing on the estimation of parameter uncertainty [18–20]. Recently, a resampling jackknife technique has been proposed to estimate standard errors in PARAFAC parameters [17]. In the present report, a noise addition Monte Carlo approach is followed in order to assess whether simple error propagation may be useful for estimating uncertainties in the concentrations predicted by the PARAFAC model. Random noise was added to calibration concentrations and to second-order signals for both theoretical and experimental examples. Noise addition has been selected because it usually gives improved results as compared to other techniques for variance estimation [9], and also because it provides better insight into the effect of real uncertainty sources.

2. Theory

2.1. The PARAFAC model

When a sample produces a $J \times K$ data matrix (a second-order array, where J and K denote the number of data points in the first and second dimension, respectively), the corresponding set obtained by 'stacking' the training matrices is a three-dimensional array. Appropriate dimensions of such an array are $I \times J \times K$ (I = number of samples). If the data follow the trilinear PARAFAC model, the array can be written as a sum of Kronecker products of three vectors for each responsive component. If $\mathbf{a}_n$, $\mathbf{b}_n$ and $\mathbf{c}_n$ collect the relative concentrations or scores ($I \times 1$), first-dimension profile ($J \times 1$) and second-dimension profile ($K \times 1$) for component n, respectively, the data array $\underline{\mathbf{X}}$ can be written as [21,22]:

$$\underline{\mathbf{X}} = \sum_{n=1}^{N} \mathbf{a}_n \otimes \mathbf{b}_n \otimes \mathbf{c}_n + \underline{\mathbf{E}} \qquad (1)$$

where $\otimes$ indicates the well-known Kronecker product, N is the total number of responsive components and $\underline{\mathbf{E}}$ is a residual error term of the same dimensions as $\underline{\mathbf{X}}$. The column vectors $\mathbf{a}_n$, $\mathbf{b}_n$ and $\mathbf{c}_n$ are usually collected into the score matrix **A** and the loading matrices **B** and **C**.

The model described by Eq. (1) defines a unique decomposition of $\underline{\mathbf{X}}$, which provides access to spectral profiles (**B** and **C**) and relative concentrations (**A**) of individual components in the I mixtures, whether they are chemically known or not. Several methods exist for the convenient analysis of three-way data, notably PARAFAC [2] and GRAM [23]. The first is guided by an alternating least-squares (ALS) minimisation (see Refs. [2,18,19]), whereas GRAM operates by directly solving an eigenvalue–eigenvector problem. From the analytical point of view, a relevant difference between these two methodologies lies in the fact that PARAFAC is able to handle multiple calibration standards, whereas GRAM usually operates with a single analyte standard.

Specific details concerning the use of PARAFAC for multiple sample calibration are given in Table 1. This scheme is useful in identifying the sources of uncertainty and their propagation to the final prediction of the analyte concentration. On one hand, uncertainties occur in the measured instrumental responses, which affect the decomposition step 2, producing **A**, **B** and **C** matrices, which will deviate from the true scores and profiles. This propagates to calibration $\mathbf{a}_n$ values, the latter being employed for the regression step 5 in Table 1. Signal errors are also contained in the $\mathbf{X}_{unk}$ data matrix for the unknown sample, and will propagate to the unknown $a_{n,unk}$ value (step 4 in Table 1). On the other hand, errors in calibration concentrations will directly affect the regression step 5 (Table 1). The purpose of the present work is to show that the standard error in the predicted concentrations can be approximated by a first-order error propagation equation, considering the pseudo-univariate PARAFAC calibration model as analogous to the well-known univariate calibration of a single analyte.

2.2. Estimation of variances

The estimation of variances in predicted concentrations including uncertainties in both concentrations and signals has already been discussed for first-order multivariate calibration methods such as CLS, PCR and PLS [10]. In these first-order methods, an expression for the prediction error

variance has been obtained by starting from the general prediction equation:

$$y_{n,\mathrm{unk}} = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_n \tag{2}$$

where $y_{n,\mathrm{unk}}$ is the predicted concentration of analyte n in the unknown sample and $\mathbf{x}$ and $\boldsymbol{\beta}_n$ are the vectors of sample instrumental response and regression coefficients for n, respectively. By performing first-order error propagation on Eq. (2), the variance in the predicted analyte concentration [var($y_{n,\mathrm{unk}}$)] is obtained as:

$$\mathrm{var}(y_{n,\mathrm{unk}}) = \mathbf{x}^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta}_n)\mathbf{x} + \boldsymbol{\beta}_n^{\mathrm{T}}\mathbf{V}(\mathbf{x})\boldsymbol{\beta}_n \tag{3}$$

where $\mathbf{V}(\boldsymbol{\beta}_n)$and $\mathbf{V}(\mathbf{x})$ are the covariance matrices of $\boldsymbol{\beta}_n$ and $\mathbf{x}$. The standard error in the prediction is, as usual, the square root of the value given by Eq. (3). The properties of the covariance terms in Eq. (3) have already been discussed in great detail [10]. Provided the errors are homoscedastic and uncorrelated, Eq. (3) can be worked out to yield [10]:

$$s(y_{n,\mathrm{unk}}) = [\mathrm{var}(y_{n,\mathrm{unk}})]^{1/2} = [hs_y^2 + (1+h)\mathrm{SEN}_n^{-2}s_x^2]^{1/2} \tag{4}$$

where $s(y_{n,\mathrm{unk}})$ is the standard error in the predicted concentration, h is the leverage, which specifies the position of the unknown sample in the calibration space [4], s_y^2 is the variance in the calibration concentrations, s_x^2 is the variance in the analytical signals and SEN_n is the analyte sensitivity, identified as [24]:

$$\mathrm{SEN}_n = \|\boldsymbol{\beta}_n\|^{-1} \tag{5}$$

where $\|\cdot\|$ indicates the Euclidean norm. The first term in Eq. (4) accounts for uncertainty in calibration concentrations, $h\mathrm{SEN}_n^{-2}s_x^2$ corresponds to uncertainties in the model parameters (regression coefficients), while the contribution of $\mathrm{SEN}_n^{-2}s_x^2$ stems from the uncertainty in the unknown sample signals. The above discussion implies no mean-centering, otherwise h has to be replaced by $h+I^{-1}$, with terms containing I^{-1} accounting for uncertainties introduced by the optional mean-centering process [10].

Eq. (4) will be probed below as regards the second-order multivariate model PARAFAC, by including appropriate values for the sensitivity and leverage parameters. This analogy is shown to provide reasonable uncertainty estimations, although the specific covariance properties of the PARAFAC regression coefficients are not available.

2.2.1. Sensitivity

A parameter, which deserves a comment in the present context, is the sensitivity. It measures the portion of the signal that is useful for prediction of a given analyte at unit concentration. Given the PARAFAC calibration model described in Table 1, one would be tempted to introduce in Eq. (4) a sensitivity value equal to the slope of the pseudo-univariate calibration graph (i.e., k in Table 1, step 5). However, the correct sensitivity factor appearing in Eq. (4) should be calculated from the vector of regression coefficients $\boldsymbol{\beta}_n$ for component n provided by the multivariate model (see Eq. (5)), which can be identified for PARAFAC from steps 4, 5 and 6 in Table 1 as:

$$\boldsymbol{\beta}_n = (\mathbf{z}_n^{+})^{\mathrm{T}}/k \tag{6}$$

where k is the slope of the calibration graph provided by the training mixtures and $\mathbf{z}_n^{+}$ is the nth row of the matrix $\mathbf{Z}^{+}$ (see Table 1), and hence:

$$\mathrm{SEN}_n = k\,\|\mathbf{z}_n^{+}\|^{-1} \tag{7}$$

This equation will be employed for computing the concentration standard errors when applying the PARAFAC model. Eq. (7) implies that the correct value of SEN_n to be used in Eq. (4) is the full sensitivity (calibration slope k), decreased by a factor which depends on the spectral overlapping ($\|\mathbf{z}_n^{+}\|^{-1}$).

2.2.2. Leverage

The leverage is an important parameter which establishes the position of an unknown sample in the calibration space [4,24]. PARAFAC is able to extract information regarding the component of interest by efficient decomposition of the array of data generated by multiple training samples. This implies that each sample leverage can be calculated by an equation similar to zero-order univariate calibration, i.e. regardless of the presence of the other components. This has been done in the framework of GRAM [15] for a single calibration standard. Employing an analogy with GRAM in which the leverage is extended to multiple standards, we suggest that the sample leverage should be given by:

$$h_{\mathrm{PARAFAC}} = y_{n,\mathrm{unk}}^2 \Big/ \sum_{i=1}^{I} y_{n,\mathrm{cal},i}^2 \tag{8}$$

where $y_{n,\mathrm{cal},i}$ is the ith calibration concentration. Therefore, estimation of the standard error of prediction on employing the PARAFAC model will be carried out using Eq. (4), by inclusion of the leverage given by Eq. (8). The adequacy of Eq. (8) is supported by the good agreement between Monte Carlo estimated uncertainties and those provided by Eq. (4).

In contrast to multivariate models based on first-order data, Eq. (8) provides a leverage value, which is independent on the presence of other sample constituents. In the partial least-squares version PLS-1, for example, the leverage depends on the concentrations of all components [24]. The specific equation, which will be used below for a comparison with results for second-order data, is [24]:

$$h_{\mathrm{PLS}} = \|\mathbf{T}^{+}\mathbf{t}_{\mathrm{unk}}\|^2 \tag{9}$$

where $\mathbf{T}$ is the matrix of PLS-1 calibration scores and $\mathbf{t}_{\mathrm{unk}}$ is the vector of scores for the unknown sample.

3. Data sets

3.1. Theoretical data sets

Two-component three-way data sets were constructed in the following way. The noiseless profiles $\mathbf{s}_{1n}$ and $\mathbf{s}_{2n}$ for each component n in the first and second dimension respectively (at unit concentration) are shown in Fig. 1. From these theoretical profiles, matrices were created with component concentrations given by a central composite design (nine samples) in which both ranges were from 0 to 1 (see the Supplementary Material for a description of the set design). Random Gaussian noise was added to calibration concentrations and responses. Specifically, calibration data matrices $\mathbf{X}_{\text{cal},i}$ were created from:

$$\mathbf{X}_{\text{cal},i} = \left[\sum_{n=1}^{2}(y_{n,\text{cal},i} + r_{ni}s_{ny})(\mathbf{s}_{1n} \otimes \mathbf{s}_{2n})\right] + \mathbf{R}s_x \qquad (10)$$

where $y_{n,\text{cal},i}$ is the nominal concentration of analyte n in the calibration standard i, s_x is the standard deviation of the noise added to signals, s_{ny} are the standard deviations for the component concentrations (in what follows, however, we will assume $s_{1y}=s_{2y}=s_y$ for simplicity, although if the concentration ranges of the analytes are different, then it is understandable that the corresponding uncertainties will also differ in range), r_{ni} is a random number (one for each analyte and for each sample) taken from a Gaussian distribution with unit standard deviation and $\mathbf{R}$ is a matrix of Gaussian random numbers of size $J \times K$.

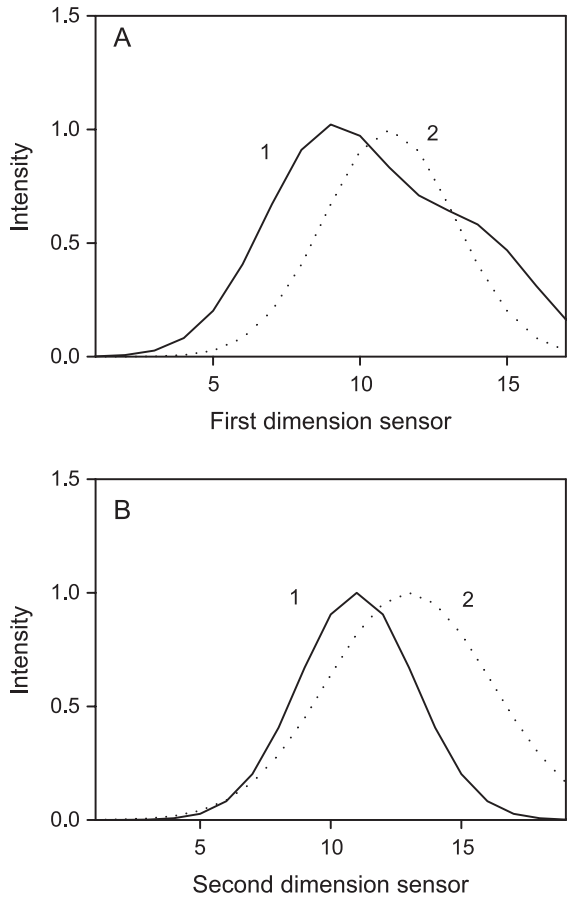

Fig. 1. Noiseless theoretical profiles for components 1 and 2 at unit concentration ($\mathbf{s}_{1n}$ and $\mathbf{s}_{2n}$), as indicated. (A) Profiles in the first dimension. (B) Profiles in the second dimension.

A set of eight additional samples (test set no. 1) with random concentrations of both components (in the range 0–1) was also created from the same profiles, according to:

$$\mathbf{X}_{\text{unk}} = \left[\sum_{n=1}^{2} y_{n,\text{unk}}(\mathbf{s}_{1n} \otimes \mathbf{s}_{2n})\right] + \mathbf{R}s_x \qquad (11)$$

where $y_{n,\text{unk}}$ is the nominal concentration of each component in the unknown. An additional set (test set no. 2) was constructed for comparison with PLS-1 (see below). It should be noticed that the unknown samples do not contain constituents, which are not modelled by the calibration set.

All nine combinations of three standard deviations for concentrations ($s_y = 0.001$, 0.01 and 0.1) and for intensities ($s_x = 0.001$, 0.01 and 0.1) were considered in building the calibration and test set no. 1. At the largest of the selected noise levels, the relative uncertainty in calibration ranges from ca. 6–20% both for total analyte content and maximum instrumental intensity. The calibration/prediction procedure was repeated for these nine combinations using different random seeds in each case (usually 10,000 times until proper convergence was achieved) and a statistic was registered of the predicted concentrations for each test sample. Finally, the 144 Monte Carlo standard errors (nine combinations of s_y and s_x, and two analytes in eight test samples) were compared to the values furnished by the simple error propagation Eq. (4), with SEN_n as in Eq. (7) and h given as h_{PARAFAC} in Eq. (8). The latter parameters were computed only once and were not averaged over the repeated calculations.

All Monte Carlo calculations were carried out with suitable MATLAB 5.3 routines [25]. Those for employing the PARAFAC model are available on the internet [26], as described in Ref. [27].

3.2. Experimental data sets

Experimental data consist of a training set of fifteen samples containing ternary aqueous mixtures of the antibiotics norfloxacin (NOR), enoxacin (ENO) and ofloxacin (OFL), with concentrations given by a central composite design. The concentration ranges are 0–25, 0–300 and 0–80 μg l^{-1}, respectively. Eight test ternary samples were prepared with random concentrations of the three analytes, all within the corresponding calibration ranges. Excitation–emission fluorescence matrices for these samples were all measured in an Aminco Bowman Series 2 spectrofluorometer, in the following spectral ranges: emission, from 378 to 501 nm each 3 nm, and excitation, 260 to 330 nm each 5

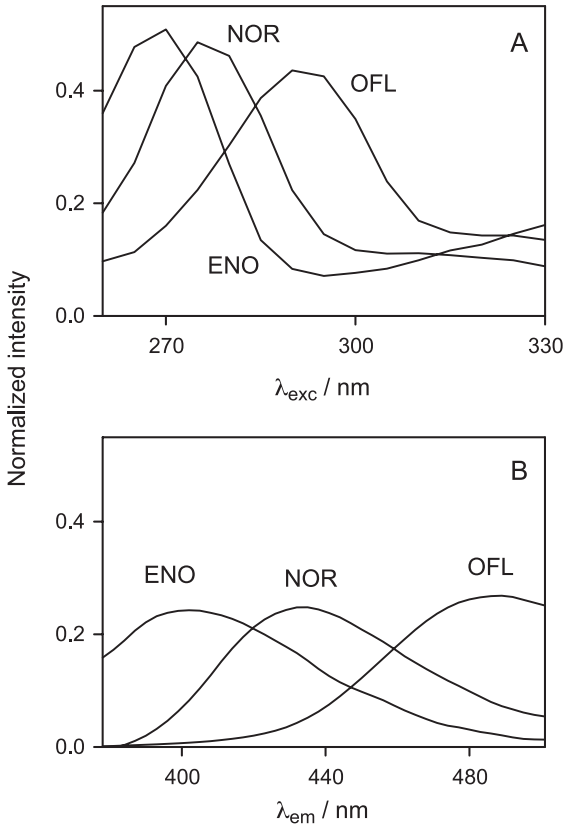

Fig. 2. (A) Normalized (to unit length) experimental excitation spectra for the three studied analytes, as indicated (at $\lambda_{em}=435$, 400 and 485 nm for NOR, ENO and OFL, respectively). (B) Normalized experimental emission spectra, using $\lambda_{exc}=277$ nm.

nm, making a total of $42\times15=630$ data points for each sample. The excitation and emission profiles for the individual components are shown in Fig. 2.

4. Results and discussion

4.1. Theoretical data sets

The theoretical data sets described in Section 3.1 were subjected to Monte Carlo repeated calculations. Specific results obtained by noise addition for one of the binary samples of the test set no. 1 are shown in Fig. 3A, which displays the 10,000 predicted concentrations for component 1 in the first test sample when using $s_y=0.01$ and $s_x=0.01$ in Eqs. (10) and (11). This simulation required Eqs. (10) and (11) to be implemented 10,000 times using in each case different random seeds, and then carrying out the steps outlined in Table 1 for prediction of analyte concentration. The statistical analysis of these results is shown in Fig. 3B in the form of a histogram, with a superimposed Gaussian function with a standard deviation of 0.0066 units. Inserting in Eq. (4) the appropriate values for component 1 in this sample gives a standard error of 0.0064, in good agreement with the simulation.

As can be seen in Fig. 3B, predicted concentrations closely follow a Gaussian distribution. Similar analyses allow one to estimate the Monte Carlo standard errors for the predicted concentrations of both components in the eight synthetic samples of test set no. 1. The values obtained after these simulations are compared with those provided by Eq. (4) in Table 2 for two selected samples of the test set no. 1, and for all nine combinations of the noise added through the different s_y and s_x values (details on the full set are given in the Supplementary Material). Fig. 4 shows that the comparison of standard errors for both predicted component concentrations in all the theoretical mixtures composing the test set no. 1 is satisfactory.

To further ascertain the independence of PARAFAC leverages on component concentrations, similar Monte Carlo calculations were carried out for a group of samples having the compositions shown in Table 3 (test set no. 2), using the above discussed calibration set. As can be seen, fifteen test samples were created in which the concentra-

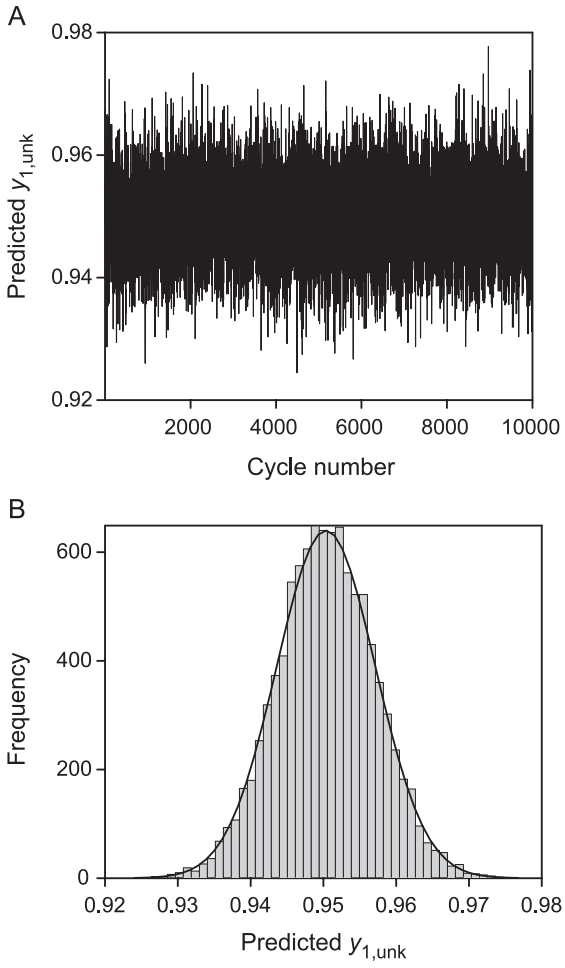

Fig. 3. (A) Predicted 10,000 Monte Carlo concentrations for analyte 1 ($y_{1,unk}$) in the first theoretical sample of test set no. 1 (using $s_y=0.01$, $s_x=0.01$ for noise addition). The nominal concentration of component 1 in this sample is 0.95. (B) Histogram showing the number of occurrences of a given predicted concentration. The solid line is a Gaussian function with a standard deviation of 0.0066 concentration units.

Table 2
Standard error in the concentrations predicted for components 1 and 2 by PARAFAC in two selected samples of the theoretical test set no. 1, as computed by Monte Carlo simulations and error propagation (Eq. (4))

Added noise[a]		Component 1			Component 2		
s_y	s_x	Nominal concentration	Standard errors		Nominal concentration	Standard errors	
			Monte Carlo	Eq. (4)		Monte Carlo	Eq. (4)
0.1	0.001	0.95	0.0532	0.0538	0.82	0.0456	0.0466
		0.45	0.0129	0.0130	0.17	0.0244	0.0250
0.01	0.001	0.95	0.0053	0.0053	0.82	0.0045	0.0046
		0.45	0.0013	0.0013	0.17	0.0024	0.0025
0.001	0.001	0.95	0.0007	0.0006	0.82	0.0006	0.0006
		0.45	0.0004	0.0004	0.17	0.0004	0.0004
0.1	0.01	0.95	0.0519	0.0504	0.82	0.0453	0.0458
		0.45	0.0130	0.0126	0.17	0.0244	0.0245
0.01	0.01	0.95	0.0066	0.0064	0.82	0.0067	0.0059
		0.45	0.0041	0.0036	0.17	0.0044	0.0042
0.001	0.01	0.95	0.0039	0.0038	0.82	0.0046	0.0037
		0.45	0.0039	0.0034	0.17	0.0035	0.0035
0.1	0.1	0.95	0.0631	0.0694	0.82	0.0669	0.0672
		0.45	0.0405	0.0392	0.17	0.0424	0.0464
0.01	0.1	0.95	0.0388	0.0365	0.82	0.0491	0.0335
		0.45	0.0395	0.0323	0.17	0.0351	0.0308
0.001	0.1	0.95	0.0399	0.0348	0.82	0.0470	0.0377
		0.45	0.0392	0.0309	0.17	0.0343	0.0357

[a] Noise addition parameters: s_y, standard deviation in calibration concentrations; s_x, standard deviation in analytical signals.

tion of analyte 1 was kept constant at three different values, whereas that for analyte 2 varied in the range 0–1 (five equally spaced values). Leverages appropriate for PARAFAC [h_{PARAFAC} in Eq. (8)] are shown in Table 3, along with the Monte Carlo standard errors and the values calculated with Eq. (4), choosing values of s_y and s_x which would highlight the leverage effects (i.e., $s_y = 0.1$, $s_x = 0.001$, for which $s_y \gg s_x$). The agreement is pleasing, indicating that the model is indeed able to separate the information concerning component 1 from that for component 2.

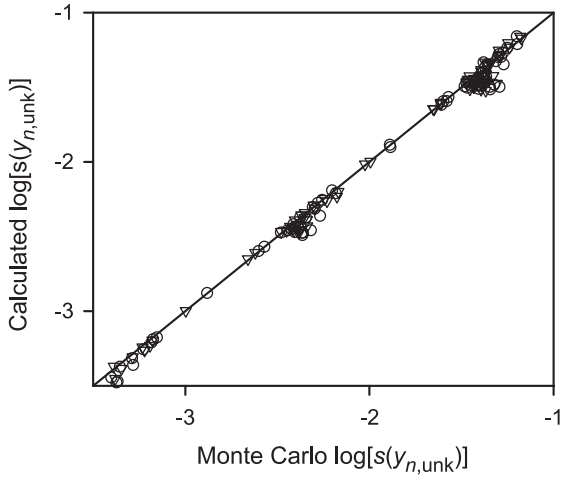

Fig. 4. Calculated (Eq. (4)) standard errors in predicted concentration $s(y_{n,\mathrm{unk}})$ versus Monte Carlo obtained values, for the eight samples of test set no. 1 and the nine combinations of s_y and s_x values (see text): component 1 (circles) and component 2 (triangles). The solid line corresponds to agreement between Monte Carlo and calculated values.

In comparison, first-order multivariate calibration performed by use of unidimensional spectra leads to leverages, which are dependent on all component concentrations. To confirm this point, PLS-1 regression analysis was carried out using the second-order data for the 9 calibration samples and the 15 theoretical samples of test set no. 2, but projected onto the second dimension. The standard errors calculated with the aid of Eq. (4) [and inserting the appropriate h_{PLS} leverage computed according to Eq. (9)] were then compared with those rendered by Monte Carlo simulations, of the type already described in Ref. [11]. Table 3 displays the specific sample leverages, which are now dependent on both component concentrations (compare with the h_{PARAFAC} values). Good agreement is found between simulations and theory (Table 3).

In summary, when using the PARAFAC model, the standard error of a predicted concentration for a given component does not depend on the concentration of other sample constituents. There is an apparent advantage in having simple leverage expressions, which are easily grasped by analogy with classical univariate calibration. This adds up to the already appreciated superiority of PARAFAC over PLS in terms of model interpretability.

The above results do also have implications regarding the estimation of the limit of detection of an analytical methodology based on PARAFAC decomposition (see Appendix A).

4.2. Experimental data sets

The application of PARAFAC error propagation equations in an experimental case requires both s_y and s_x to be known. The standard error s_y is usually available to expe-

Table 3
Concentrations of components 1 and 2, leverages and standard errors in concentrations of analyte 1 predicted by PARAFAC and PLS-1 in test set no. 2, as computed by Monte Carlo simulation and error propagation (Eq. (4))[a]

Test sample	Nominal concentrations		PARAFAC[b]			PLS-1[c]		
	Component		Leverage	Standard errors		Leverage	Standard errors	
	1	2	$h_{PARAFAC}$	Monte Carlo	Eq. (4)	h_{PLS}	Monte Carlo	Eq. (4)
1	0.1	0.2	0.003	0.006	0.006	0.013	0.011	0.012
2	0.1	0.4	0.003	0.006	0.006	0.068	0.026	0.026
3	0.1	0.6	0.003	0.006	0.006	0.171	0.041	0.039
4	0.1	0.8	0.003	0.006	0.006	0.321	0.057	0.056
5	0.1	1.0	0.003	0.006	0.006	0.519	0.072	0.073
6	0.5	0.2	0.077	0.027	0.029	0.090	0.030	0.030
7	0.5	0.4	0.077	0.027	0.029	0.079	0.028	0.028
8	0.5	0.6	0.077	0.027	0.029	0.115	0.035	0.037
9	0.5	0.8	0.077	0.027	0.029	0.199	0.045	0.044
10	0.5	1.0	0.077	0.027	0.029	0.331	0.057	0.057
11	0.9	0.2	0.250	0.050	0.053	0.358	0.060	0.061
12	0.9	0.4	0.250	0.050	0.053	0.280	0.053	0.051
13	0.9	0.6	0.250	0.050	0.053	0.250	0.050	0.052
14	0.9	0.8	0.250	0.050	0.053	0.268	0.052	0.052
15	0.9	1.0	0.250	0.050	0.053	0.334	0.057	0.059

[a] Noise addition parameters: $s_y = 0.1$ (standard deviation in calibration concentrations), $s_x = 0.001$ (standard deviation in analytical signals).
[b] The leverage $h_{PARAFAC}$ was computed using Eq. (8).
[c] The leverage h_{PLS} was computed using Eq. (9).

rienced users from the uncertainty propagation in preparing standard samples, or, if a reference method is employed to determine them, from its estimated standard error. In the experimental samples under study, a value of $s_y = 0.05$ μg l^{-1} has been estimated. As regards s_x, the level of instrumental noise was approximated by spectral replicate measurements as ca. 0.1 arbitrary fluorescence units, although it can also be estimated from the fitting residuals of the PARAFAC model. Shown in Fig. 5 is a comparison between standard errors calculated from error propagation and those estimated by Monte Carlo noise addition to both calibration concentration and to training and unknown spectra. The agreement can be viewed as reasonable, and lends additional support for the use of the approximate variance equations discussed in the present work.

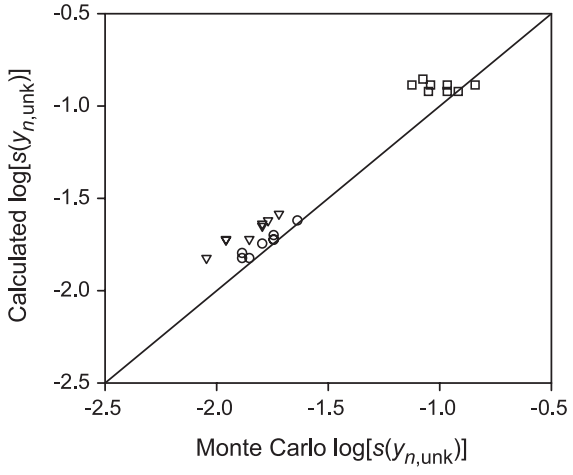

Fig. 5. Calculated (Eq. (4)) standard errors in predicted concentration $s(y_{n,unk})$ versus Monte Carlo obtained values, for the eight experimental test samples (see text): ENO (squares), NOR (circles) and OFL (triangles). The solid line corresponds to agreement between Monte Carlo and calculated values.

5. Conclusions

Monte Carlo simulations based on noise addition suggest that a simple error propagation approach is useful for estimating sample-specific standard errors when concentrations are predicted by a three-way PARAFAC model. The employed expression enables one to set up prediction intervals, which is better than relying on mean prediction errors for a group of test samples. The limit of detection is shown to be component-specific, an advantage over other multivariate methods in terms of interpretability. Since predictions should always be accompanied by realistic uncertainties, it is advisable to be aware of the errors in instrumental measurements and also in reference concentration values before applying the variance equations. Further work is in progress to cases where PARAFAC exploits the second-order advantage by including data for the unknown into the array before decomposition.

Acknowledgements

A.C.O. thanks the University of Rosario, CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas, Project PIP 431), ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica, Project PICT 99 No. 06-06078) and the John Simon Guggenheim Foundation for

a fellowship. The authors wish to express their gratitude to the reviewers for helping to improve the presentation.

Appendix A. Implications for limit of detection estimation

The above mentioned properties of PARAFAC have important implications regarding the limit of detection (LOD) of the analytical method. A rigorous definition of the latter is [3,14]:

$$\mathrm{LOD} = z_{1-\alpha}s(0) + z_{1-\beta}s(D) \quad \text{(A1)}$$

where $s(0)$ is the standard error in the predicted concentration for a blank sample, i.e., a sample not containing the analyte of interest, $s(D)$ is the standard error for a sample where the analyte concentration is equal to LOD and $z_{1-\alpha}$ and $z_{1-\beta}$ are upper α and β percentage points of a normal distribution, respectively. As discussed above, when using PARAFAC to model three-way data, $s(0)$ and $s(D)$ would be given by:

$$s(0) = [h_0 s_y^2 + (1 + h_0)\mathrm{SEN}^{-2}s_x^2]^{1/2} \quad \text{(A2)}$$

$$s(D) = [h_D s_y^2 + (1 + h_D)\mathrm{SEN}^{-2}s_x^2]^{1/2} \quad \text{(A3)}$$

where h_0 h_D are calculated by inserting and $y_{n,\mathrm{unk}} = 0$ and $y_{n,\mathrm{unk}} = \mathrm{LOD}$ in Eq. (8) (no mean centering is implied). The LOD is thus obtained by combining Eqs. (A1), (A2) and (A3). In the event that $h_\mathrm{D} \ll 1$, the value is:

$$\mathrm{LOD} = 3.3s(0) \quad \text{(A4)}$$

where $z_{1-\alpha}$ and $z_{1-\beta}$ are at 95% confidence level.

Since $s(0)$ is analyte-specific, the LOD for a given component is seen to be independent on the concentrations of other analytes. This is in contrast to first-order multivariate calibration using PLS (or other techniques such as CLS, PCR, etc.) in which LODs for specific analytes cannot be defined in isolation.

References

[1] K.S. Booksh, B.R. Kowalski, Anal. Chem. 66 (1994) 782A–791A.
[2] R. Bro, Chemom. Intell. Lab. Syst. 38 (1997) 149–171.
[3] L.A. Currie, Anal. Chim. Acta 391 (1999) 105–126.
[4] H. Martens, T. Næs, Multivariate Calibration, 2nd ed., Wiley, Chichester, UK, 1989.
[5] A. Maroto, J. Riu, R. Boqué, X. Rius, Anal. Chim. Acta 391 (1999) 173–185.
[6] B. Efron, G. Gong, Am. Stat. 37 (1983) 36–48.
[7] C.N. Léger, D.N. Politis, J.P. Romano, Technometrics 34 (1992) 378–399.
[8] H.R. Keller, J. Röttele, H. Bartels, Anal. Chem. 66 (1994) 937–943.
[9] N.M. Faber, Chemom. Intell. Lab. Syst. 64 (2002) 169–179.
[10] K. Faber, B.R. Kowalski, J. Chemom. 11 (1997) 181–238.
[11] A.C. Olivieri, J. Chemom. 16 (2002) 207–217.
[12] G. Baffi, E. Martin, J. Morris, Chemom. Intell. Lab. Syst. 61 (2002) 151–165.
[13] M. Linder, R. Sundberg, J. Chemom. 16 (2002) 12–27.
[14] N.M. Faber, R. Boqué, J. Ferré, Chemom. Intell. Lab. Syst. 55 (2001) 91–100.
[15] R. Boqué, J. Ferré, N.M. Faber, F.X. Rius, Anal. Chim. Acta 451 (2002) 313–321.
[16] N.M. Faber, R. Bro, Chemom. Intell. Lab. Syst. 61 (2002) 133–149.
[17] J. Riu, R. Bro, Chemom. Intell. Lab. Syst. 65 (2003) 35–49.
[18] P. Paatero, Chemom. Intell. Lab. Syst. 38 (1997) 223–242.
[19] X. Liu, N.D. Sidiropoulos, IEEE Transactions on Signal Processing 49 (2001) 2074–2086.
[20] P. Paatero, J. Comput. Graph. Stat. 8 (1999) 854–888.
[21] G.W. Ewing, Instrumental Methods of Chemical Analysis, McGraw-Hill, New York, 1985.
[22] S. Leurgans, R.T. Ross, Stat. Sci. 7 (1992) 289–319.
[23] E. Sánchez, B.R. Kowalski, Anal. Chem. 58 (1986) 496–499.
[24] A. Lorber, A. Harel, Z. Goldbart, I.B. Brenner, Anal. Chem. 59 (1987) 1260–1266.
[25] MATLAB 5.3, The MathWorks, Natick, Massachusetts, USA, 1999.
[26] http://www.models.kvl.dk/source/.
[27] C.A. Andersson, R. Bro, Chemom. Intell. Lab. Syst. 52 (2000) 1–4.
[28] C.R. Rao, S. Mitra, Generalized Inverse of Matrices and its Applications, Wiley, New York, 1971.